Data and Metadata Reporting Standards for the U.S. Environmental Protection Agency's PM Supersites Research Program[1]

Les A. Hook, NARSTO Quality Systems Science Center, Oak Ridge National Laboratory[2]
Sigurd W. Christensen, NARSTO Quality Systems Science Center, Oak Ridge National Laboratory
William B. Sukloff, Environment Canada, Meteorological Service of Canada

Abstract - The EPA Supersites Research Program needs consistency of metadata and data structures to facilitate information sharing among investigators, analysts, and ultimately secondary data users. Under the auspices of NARSTO[3] a successful mechanism was created to develop and implement reporting standards. The development effort included working closely with Supersites data coordinators, investigators, and technical experts, and also leveraging from existing data standards and practices. Overall, the standards are getting good acceptance from the atmospheric research community.

The U.S. Environmental Protection Agency is sponsoring a major atmospheric particulate matter (PM) data collection effort in seven major U.S. cities, called the PM Supersites Research Program (Fig. 1). The Supersites Program's objectives are to (1) characterize PM and its constituents, (2) collect data and samples to support health and exposure studies to reduce uncertainty in setting National Ambient Air Quality Standards, and (3) compare emerging sampling and analysis methods with routine techniques to enable a smooth transition to advanced methods. In addition to analyzing individual site PM and atmospheric conditions, the data from all the Supersites are to be capable of being integrated for cross-site analyses, and are to be archived in a timely manner and be readily available to the public.

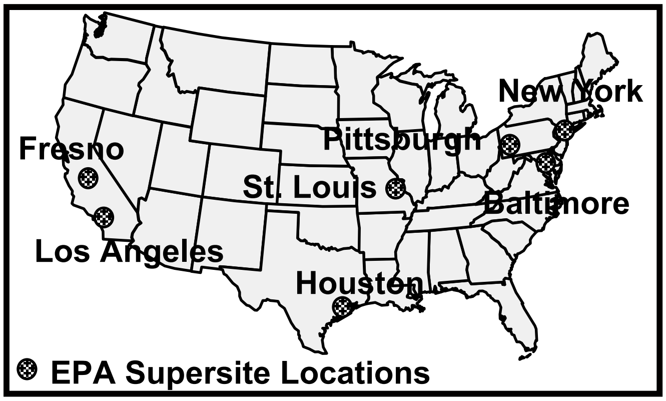

Figure 1. U.S. EPA Supersites Research Program

Data reporting was addressed in the Cooperative Agreements that implement the Program. Data Coordinators support the data reporting process at each Supersite (Fig. 2). The NARSTO Permanent Data Archive (PDA) at the Langley NASA DAAC was designated the final repository. The PDA has a required self-documenting data format, the NARSTO Data Exchange Standard (DES), with several metadata requirements. The archiving process is mediated by the NARSTO Quality Systems Science Center (QSSC), Oak Ridge National Laboratory. NARSTO encourages scientists to document their data at a level sufficient to satisfy the well-known "20-year test". That is, someone 20 years from now, not familiar with the data or how they were obtained, should be able to find data of interest and then fully understand and use the data solely with the aid of the

[1] Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency. Michael Jones, Project Officer.

[2] ORNL research was sponsored by U.S Environmental Protection Agency and U.S. Department of Energy and performed at Oak Ridge National Laboratory (ORNL). ORNL is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

[3] NARSTO is a tri-national, public-private partnership for dealing with multiple features of tropospheric pollution, including ozone and suspended particulate matter.

documentation archived with the data[4]. Integration of data in future analyses demands consistently defined metadata elements and values.

Data and Metadata Standards Development

We began the development of the data reporting standards began by working with the Data Management Coordinators for each Supersite. A Data Management Working Group (DMWG) was formed with the QSSC as the lead. The Working Group communicated through weekly teleconferences to deal with consistency of metadata content and data reporting format. Minutes of the teleconferences discussion and decisions were distributed to the DMWG and Site Principal Investigators. The continued support of the EPA Program Mangers is critical to the success of this effort.

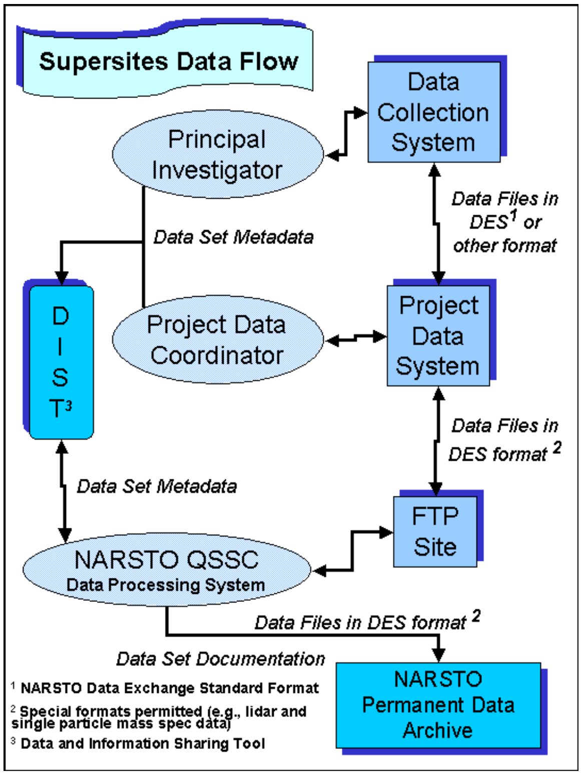
Supersites Data Flow
Principal Investigator
Data Collection System
Data Files in DES¹ or other format
Data Set Metadata
D I S T³
Project Data Coordinator
Project Data System
Data Files in DES format ²
Data Set Metadata
FTP Site
NARSTO QSSC
Data Processing System
Data Files in DES format ²
Data Set Documentation
NARSTO Permanent Data Archive
¹ NARSTO Data Exchange Standard Format
² Special formats permitted (e.g., lidar and single particle mass spec data)
³ Data and Information Sharing Tool

Figure 2. Supersites Data Flow Diagram.

We incorporated metadata elements and values from other metadata standards when available to promote consistency within EPA and the atmospheric research community, and to anticipate integrating data from additional sources. For example, in addition to existing NARSTO standards, we used site descriptors and event flags from EPA AIRS, detailed flags from EPA Region 5, use of the CAS Registry Number and CAS Index Name for chemical identification from EPA CRS, and the non-chemical variable naming syntax from the DOE ARM Program.

By leveraging existing resources and the developing data management resources and technical expertise of the individual Supersites and NARSTO, we were able to develop a set of robust reference materials and supporting systems. Site-specific implementation flexibility is always a consideration and was maintained when possible. Each metadata standard was completed within the DMWG and then sent to the Site Principal Investigators for approval, after which they were considered Consensus Metadata Standards. The DMWG updates these as needed.

Consensus Metadata Standards

Site Identification: Identifies a standard syntax for naming fixed and mobile sites used by studies or networks for air quality sampling and monitoring. A site is assigned a 12-character site identifier that includes a four-character site abbreviation (the "site mnemonic"). The first four characters identify a study or network. A master list of site names from the Supersites program will be assembled by the QSSC. The master list would include additional information about each site, as available: latitude, longitude, elevation, EPA AIRS identifier, land use, and location type.

[4] National Research Council, Committee on Geophysical Data, *Solving the Global Change Puzzle, A U.S. Strategy for Managing Data and Information,* National Academy Press, Washington, D.C., 1991.

Identifying Chemical and Physical Variables and Descriptive Metadata:

- Identifying Chemical Substances with a CAS Registry Number: Valid values are the CAS Number (with "C" prefix -- "C" prevents spreadsheet programs from converting some CAS numbers to dates) and Chemical Name (either CAS-9CI, IUPAC, or other common name). CAS numbers and preferred names for common atmospheric constituents are in a reference table.
- Identifying Chemical Substances, Calculated Quantities, and Physical/Non-chemical Measurements that do not have a designated CAS Registry Number: Variable names are formed beginning with the root concept, and followed by a detailed modifier if needed, separated by a ":". For example, PM10: area, PM10: count, PM10: mass, and Temperature: air, Temperature: dew point, Temperature: virtual. These variables can be method specific and require special differentiation. Definition of new variables is relatively straightforward when the format is followed.
- Identifying Metadata Elements: Valid variable names for metadata elements including site information, locations, dates, times, and sampling conditions are provided in a reference table. The correct format is the root concept, followed by a detailed modifier if needed, separated by a ":". For example, Date start: local time, Date end: local time, and Latitude: decimal degrees, Longitude: decimal degrees.

Data Quality Flags: Reported data values must be assigned at least one data quality flag by the data originator to indicate whether the data are valid without qualification, valid but qualified/suspect, or invalid due to serious sampling or analysis problems. These flags may be the NARSTO data qualification flags or other more detailed flags as defined by a Project. Project-defined flags must be mapped to NARSTO flags. Reference tables of NARSTO standard flags, detailed project flags, and EPA AIRS exceptional-event flags are provided for users.

Changes and additions to the reference tables are controlled. Site investigators and data users are encouraged to work with their Data Management Coordinators to suggest improvements in and additions to the reference tables. A Data Coordinator should recommend additions or changes to the DMWG and QSSC for discussion and consensus. Subject matter experts are consulted when appropriate.

Data Exchange Standard Development

Data files submitted for archiving should be in the NARSTO Data Exchange Standard (DES) format. The DES format follows a spreadsheet-compatible layout and is stored as ASCII comma-separated value (.csv) files. The DES does not rely on row position to identify metadata information, but uses tags to describe the information contained in the row. The DES is a self-documenting format with three sections: the header section contains information about the contents of the file and the data originator; the middle section contains metadata tables that describe/define sites, flags, and other codified fields; and the final section is the main data table that contains key sampling and analysis information and the data values.

The consensus metadata standards for site names, data quality flags, and parameter names, plus key characteristics (see below) are implemented in the DES. An Excel/97® template for inputting data and metadata has been developed to support data providers. The template is annotated with comments, instructions, frequently asked questions, and examples of completed files. Within the template are picklists for selecting values for various metadata fields to promote consistency in terminology.

Supersite Enhancements to the Data Exchange Standard: Until the initiation of the PM Supersites Program, the DES had been used primarily with gaseous atmospheric constituents and meteorological measurements (e.g., ozone, air temperature, and solar radiation). The sampling and measurement of these constituents is generally straightforward with well-defined methods and reporting conventions. It soon became clear that PM measurements are not so easily characterized. PM results (e.g., size-differentiated mass, number, and chemical composition) need more metadata than just the name, units, and analysis method. In many cases, results are operationally defined by the specific field sampling configurations, measurement devices, and conditions, and the laboratory sample preparation and analysis methods.

To address the expansion of measurement types, a set of key characteristics (Table 1) was defined to capture enough of the measurement information to be meaningful and helpful in a data file, while avoiding excessive detail. The key characteristics are metadata fields that hold general descriptions of the field, instrument, and laboratory conditions. Detailed information would always be included as companion files, such as the Quality Assurance Project Plans. Key characteristics, metadata values, and organization of the DES were defined through invaluable interactions of Data Coordinators, with PIs and with other field and laboratory technical experts.

Key Characteristics provide general sampling and analysis information that describes the data.

> OBSERVATION TYPE
> SAMPLING HEIGHT (M AGL)
> FIELD SAMPLING OR MEASUREMENT PRINCIPLE
> INLET TYPE
> MEDIUM
> COATING OR ABSORBING SOLUTION/MEDIA
> SAMPLING HUMIDITY OR TEMPERATURE CONTROL

> PARTICLE DIAMETER--LOWER BOUND (UM)
> PARTICLE DIAMETER--UPPER BOUND (UM)
> PARTICLE DIAMETER--MEDIAN (UM)

> WAVELENGTH (NM)
> WAVELENGTH--LOWER BOUND (NM)
> WAVELENGTH--UPPER BOUND (NM)

> SAMPLE PREPARATION
> LABORATORY ANALYTICAL METHOD

> VOLUME STANDARDIZATION
> BLANK CORRECTION

> INSTRUMENT NAME AND MODEL NUMBER
> MEASUREMENT PRINCIPAL INVESTIGATOR

> EXPLANATION OF ZERO OR NEGATIVE VALUES

> EXPLANATION OF REPORTED DETECTION LIMIT VALUES
> DETECTION LIMIT VALUES

> EXPLANATION OF REPORTED UNCERTAINTY VALUES
> UNCERTAINTY VALUES

Picklists for selecting Key Characteristic values are included in the DES template.

Table 1. Key Characteristics Included in the Data Exchange Standard.

Additional Data Reporting Guidance

To ensure that data can be integrated for successive analyses, consistently reported data and metadata are essential. Supersites' Technical and Quality Assurance Leads provided this guidance.

Submittal of Uncertainty Estimates: EPA is strongly recommending that within each Supersite the research investigators and data managers estimate and report the data uncertainties. Estimating the uncertainty of the data collected is of paramount importance to the purpose of the Supersites Project. Data users will need to understand the uncertainty of the data, which will enhance confidence in their assumptions and predictions. The DES has been updated to ensure that uncertainty can be conveniently reported by data providers and can be interpreted by data users.

Data Reporting Conventions Guidance: To further promote consistency among data products, the "level" of data to report has been specified for data providers. Mass/volume measurements (e.g., from filters) should be reported as concentrations, rather than separately as mass and volume. PM mass data should be referenced to local ambient temperature and pressure conditions to be comparable to federal reference method PM data. Associated meteorological data, including temperature and pressure conditions, should be reported either in the same file or in a referenced meteorological data file.

Similarly, units for chemical variables and particle measurements are specified to follow SI standards, when possible, or units commonly used by the research community. A pick list of units has been implemented in the DES.

Data Archiving Process Guidance

The Supersites Data Managers are provided with specific guidance for carrying out the final steps in the data management and archiving process. Specifications for data set and data file naming and configuration control are given. The NARSTO Data and Information Sharing Tool has a convenient metadata entry/export feature for efficient preparation of archive documentation. The QSSC is the source for information and assistance, and it verifies submitted DES data file format compliance and mediates these archiving activities.

Read and Verify Program Verifies Data File DES Format Compliance: Data Coordinators send the completed DES files to the QSSC. A Read and Verify Program checks numerous format and content elements of the DES files by verifying that key characteristic values are in the picklist reference tables, key phrases are correctly formed, variable formats and format types are correct, CAS numbers are in reference table, dates and times are properly formatted, the UTC time offset is correct, flags are in the flag look-up table, and sites appearing in the main data table have corresponding entries in the site information table. The Read and Verify Program also calculates and inserts into the file a set of summary statistics records for each numeric variable in the file. The statistics include minimum, maximum, mean, standard deviation, n, number of missing values, and total number of records. As an additional QA check, time series plots can be created. The plots include summary statistics and key characteristic values. (Fig. 3).

If the Read and Verify Program finds problems, a QA problem report is sent back to the data originator, with some guidance on how to correct the problems. The data originator should correct the problems and resubmit the data file. This process continues until all parties are satisfied with the dataset. Supersite Data Coordinators with SAS® software are also running this program before submitting files to the QSSC.

Data and Information Sharing Tool (DIST) Generates Archive Documentation: The NARSTO DIST was implemented for Supersites to support compiling data set metadata and generating archive documentation. Either a Data Provider or the QSSC can enter metadata into the DIST

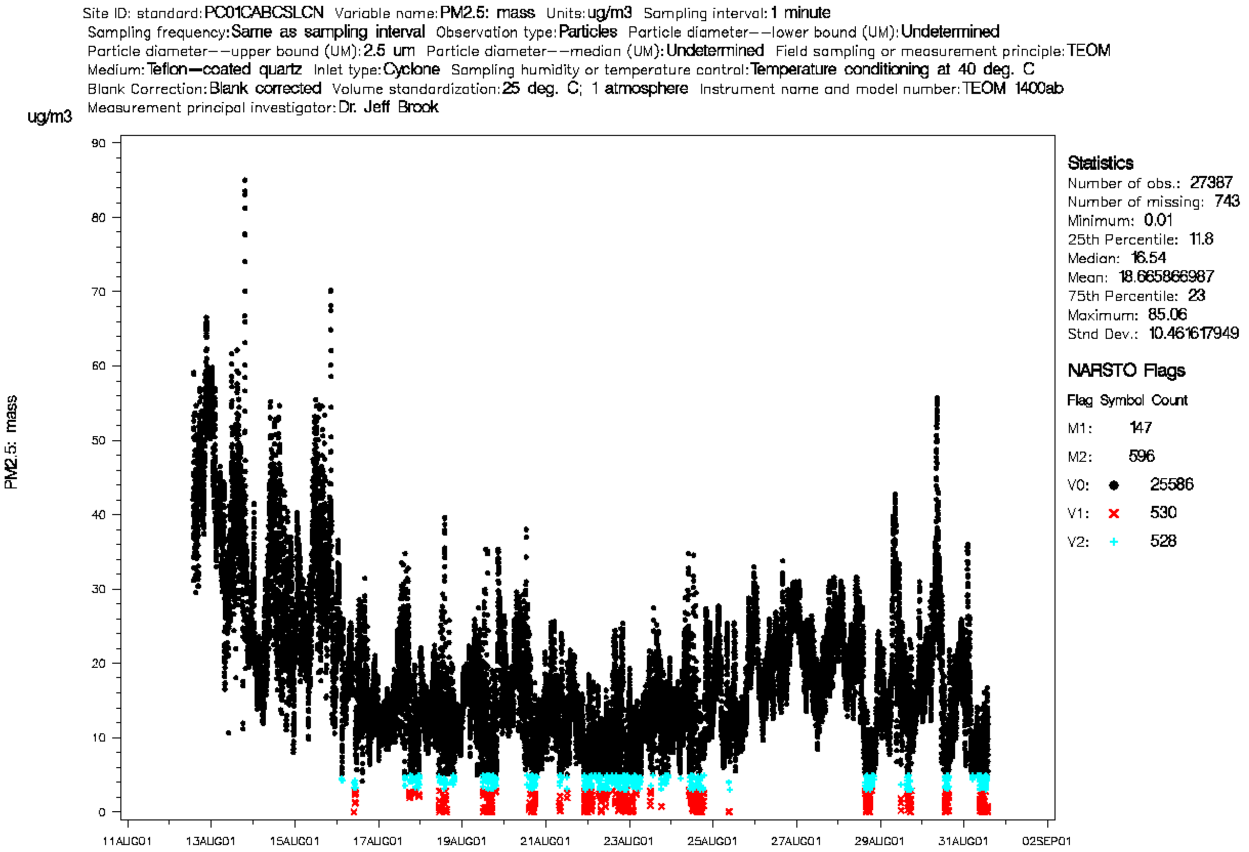

Figure 3. Example Time Series Plot Generated by Read and Verify Quality Assurance Program.

metadata editor and output it in the formats needed by the NARSTO Permanent Data Archive. When documentation is complete and the QSSC data file verification checks are complete, the properly formatted data files can be moved to the archive.

All of the standards, the DES template, DIST, and guidance documents referenced in this paper can be accessed through the QSSC web site [http://cdiac.esd.ornl.gov/programs/NARSTO/].

Conclusions

The accomplishments of this development effort to date are, in no small part, due to the early recognition of the need for data management planning and implementation and its inclusion in the Cooperative Agreements, and the continued support of the EPA Program Mangers. This standards development process was successful in integrating the input of Supersites data management and research staff with existing NARSTO and other applicable standards. The product is a robust set of Supersites data and metadata reporting standards that will facilitate current PM data reporting, analyses, and archiving activities: can be extended to additional data types; and will support integrated analyses and future research projects.

·The submitted manuscript has been authored by a contractor of the U.S. Government under contract DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.